Identifying Speaking and Drinking Events Within Audio Recordings for Multiactivity Analysis

Author: Dorothy Zhang
d.l.zhang@student.tudelft.nl
Responsible Professors: Koen Langendoen, Hayley Hung
Supervisors: Vivian Dsouza, Stephanie Tan

1. Introduction

- Multiactivity: multitasking in a social context [1]
- Reveal hidden rules of human social behaviour [1]
- When to drink and when to speak?
- Use audio to identify these actions

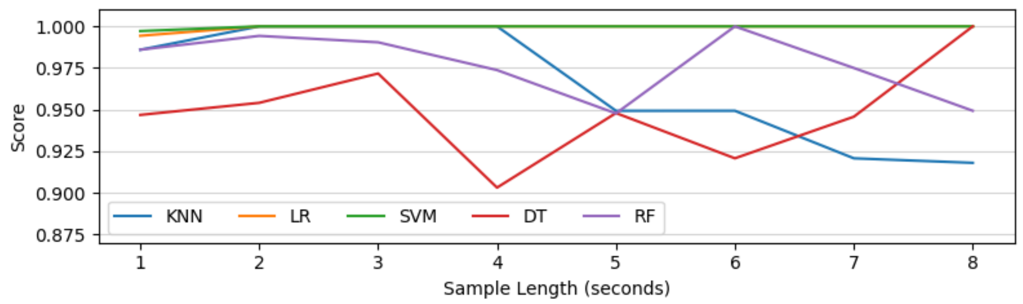
"How feasible is it to use audio recordings captured from a drinking glass to identify speaking and drinking events in social interactions?"

2. Method

1. Record audio from a drinking glass
   - Speaking
   - Drinking
   - Ambient noise
2. Extract audio features [2]
   - Mel-Frequency Cepstral Coefficients
   - Spectral (Centroid, Bandwidth, Contrast, Roll-Off)
   - Zero Crossing Rate & Root Mean Squared Energy
3. Compare different Machine Learning models [3]
   - K-Nearest Neighbours
   - Linear Regression
   - Support Vector Machine
   - Decision Tree
   - Random Forest
4. Simulate noisy environments [4]
   - Music
   - Noisy room
   - Podcast: simulating other speakers

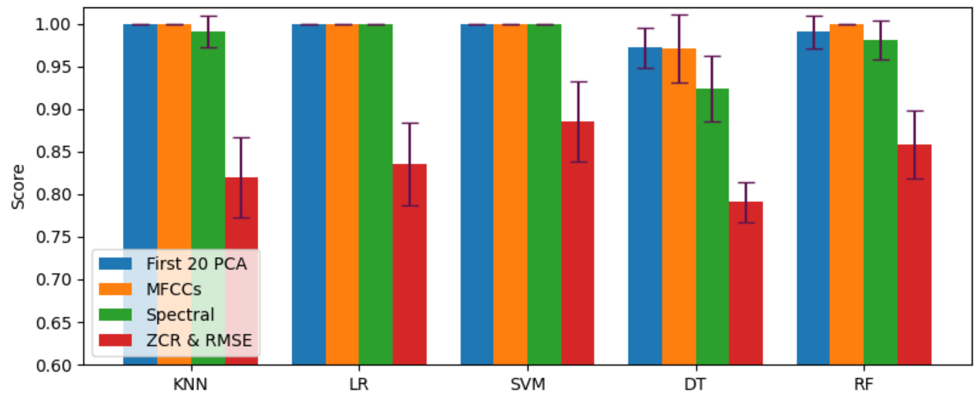
3. Results

Audio Classification Performance: Training Sample Length (F_1 score)



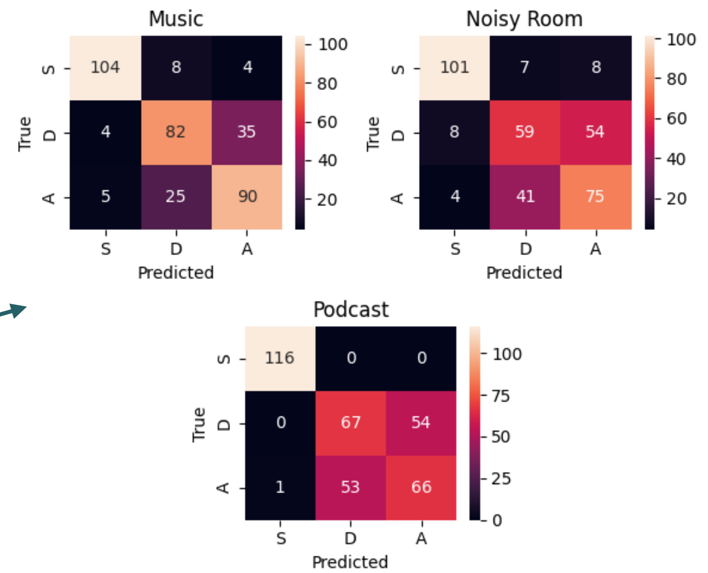
- Linear classifiers: 100% accuracy after 2-second window
- Non-linear classifiers: Fluctuate between 90% and 100%

Audio Classification Performance: Extracted Audio Features (F_1 score)



- Best: MFCCs – Average 99.4%, Spectral – Average 97.9%
- Worse: ZCR & RMSE – Average 83.8%

Audio Classification Performance:
Noisy Environments (SVM Confusion Matrices)



- Speech can still be reliably identified
- Drinking becomes less distinguishable from background noise
- Music has a lesser negative effect than a noisy room or the presence of other speakers

4. Limitations

- Small sample size
- Audio recorded of one person only
- ML model parameters not finetuned
- Noisy environments were simulated and not collected from real-life
- Use of audio alone loses information on gestures, facial expressions, etc., information within audio data limited for further in-depth analysis

5. Future Work

- Use inertial sensors to detect drinking action in noisy environments
- More diverse audio sources
- Continuous activity recognition over longer audio recordings

6. Conclusion

- Clean audio data can reach 100% classification accuracy
- Noisy environments less accurate, drinking audio more obscured
- MFCC features perform the best
- Linear ML classifiers more stable

References:
[1] P. Haddington, T. Keisanen, L. Mondada, and M. Nevile, Multiactivity in Social Interaction: Beyond multitasking. John Benjamins Publishing Company, 2014
[2] B. McFee et al., "librosa/librosa: 0.10.2.post1". Zenodo, May 14, 2024. doi: 10.5281/zenodo.11192913.
[3] F. Pedregosa et al., 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[4] I. Jordal, https://iver56.github.io/audiomentations/

TU Delft